NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

AEC Computing Facility

AEC RESEARCH AND DEVELOPMENT REPORT

PHYSICS AND MATHEMATICS NYO-7974

DIFFERENTIAL EQUATIONS, DIFFERENCE EQUATIONS AND MATRIX THEORY

by

P. D. Lax

November 15, 1957

institute of mathematical sciences

NEW YORK UNIVERSITY

NEW YORK, NEW YORK

NYO-7974

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission to the extent that such employee or contractor prepares, handles or distributes, or provides access to, any information pursuant to his employment or contract with the Commission.

AEC Computing and Applied Mathematics Center
Institute of Mathematical Sciences
New York University

PHYSICS AND MATHEMATICS

NYO-7974

DIFFERENTIAL EQUATIONS, DIFFERENCE EQUATIONS AND MATRIX THEORY

by

P. D. Lax

November 15, 1957

Contract No. AT(30-1)-1480

ABSTRACT

Friedrichs, in his studies of difference approximations to solutions of symmetric hyperbolic systems, has shown that difference schemes with positive coefficients are stable. In this paper we show that this stability criterion of Friedrichs is valid for difference approximations to arbitrary hyperbolic systems. The proof relies on the convexity and monotonicity of eigenvalues as matrix functions over a linear space of matrices with only real eigenvalues. These theorems, well known in the symmetric case, are proved with the aid of theorems concerning the dependence of solutions of hyperbolic equations on their initial data.

TABLE OF CONTENTS

Page

Abstract 2

Section

1. Introduction 4
2. Hyperbolic Equations 7
3. Some Matrix Theorems 19
4. Finite Difference Schemes 25

Bibliography 39

DIFFERENTIAL EQUATIONS, DIFFERENCE EQUATIONS AND MATRIX THEORY*†

1. Introduction.

In this paper we shall study sets of real matrices all of whose linear combinations with real coefficients have only real eigenvalues. We shall show that over a linear space spanned by such matrices, the largest eigenvalue is a convex function, the smallest is concave, and all eigenvalues are monotonic with respect to the partial ordering induced by the convex cone of matrices with positive eigenvalues. This is a generalization of classical results known for symmetric matrices - a special but important subset of our matrices. Matrices of this kind occur as coefficients of first order vectorial hyperbolic equations. The proofs of the matrix theorems mentioned before, as well as some related ones, are based on known theorems in the theory of hyperbolic equations and will be presented in section 3. The relevant portions of the theory of hyperbolic equations are sketched in section 2.

---

* This work was carried out under Contract AT(30-1)-1480 with the United States Atomic Energy Commission.

† Presented to the American Mathematical Society in Rochester, December 1956 and Washington, October 1957.

In section 4 we apply one of the matrix theorems of section 3 to derive a criterion for the stability of difference schemes used to approximate solutions of differential equations. This criterion is an extension from the symmetric to the nonsymmetric case of Friedrichs' stability theorem for difference equations with positive coefficients. Applications to various specific difference schemes are given at the end of the section.

We close this section by constructing a pair of matrices with the aforementioned property yet which cannot be simultaneously made symmetric by a similarity transformation. Our example will result from perturbing the pair of symmetric matrices

$$A_o = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} , \quad B_o = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} .$$

Every linear combination $aA_o + bB_o$ with a and b real and $a^2 + b^2 = 1$ has the eigenvalues -1, 0 and 1. Since the set of matrices with real and distinct eigenvalues is open, it follows that if X is any (real) matrix, then for ε small enough all linear combinations of A_o and B, where $B = B_o + \varepsilon X$, also have real and distinct eigenvalues. We shall show now that with an appropriate choice for X, A_o and B cannot be made symmetric by the same similarity transformation. Suppose to the contrary that there exists a matrix S such that both $S^{-1}A_0S$ and $S^{-1}BS$ are symmetric. Let U be the unitary matrix which diagonalizes $S^{-1}A_0S$; denote SU by T and $T^{-1}A_0T$ by A. Both A_0 and A are diagonal and, being similar, have the same eigenvalues. Therefore $A = A_0$(1):

$$T^{-1}A_0T = A_0 \quad .$$

Multiplying by T we get $A_0T = TA_0$, i.e. T and A_0 commute. Therefore T must have the same eigenvectors as A_0, i.e. T is diagonal. On the other hand $T^{-1}BT = U^{-1}S^{-1}BSU$ is also symmetric. Since T is diagonal, this implies, it is easy to check, that

$$b_{21}b_{32}b_{13} = b_{12}b_{23}b_{31} \quad .$$

But we can easily choose X so that this condition is not satisfied.

---

1 Strictly speaking, the diagonal elements in A and A_0 may not appear in the same order; but A can always be brought into A_0 by a unitary transformation.

2. *Hyperbolic Equations.*

We shall deal with first order systems of partial differential equations of the form

$$u_t = \sum_{k=1}^{m} A^k u_{x^k} = A \cdot u_x \tag{2.1}$$

where u is a vector of n unknown functions, and A^k are real constant coefficient matrices. The initial value problem is to find a solution of this equation with prescribed values at $t = 0$:

$$u(x,0) = \phi(x) \quad . \tag{2.2}$$

The initial value problem is said to be *properly posed* if (2.1), (2.2) has a unique solution for all sufficiently -- say k times -- differentiable data $\phi(x)$.

We start with the following classical result:

Theorem 2.1: *The initial value problem* (2.1), (2.2) *is properly posed if and only if all linear combinations* $A \cdot \xi$ *of the coefficients* A^k *with real coefficients* ξ_k *have only real eigenvalues. The solution depends continuously on the data,* i.e. *the value of the solution at any point is a continuous linear functional of the data in the* C^N-topology *for* N *large enough.*

An equation whose coefficients satisfy this condition is called hyperbolic.

The necessity of this condition has been demonstrated by Hadamard [8] and Banach [1], page 44-7. Their arguments have been extended by the present author to equations with variable coefficients, [16].

The solution can be constructed either by Fourier synthesis or by the method of plane waves (see Fritz John, [13], or Courant-Lax, [3]). Since the proof is usually carried out under slightly more stringent conditions, we sketch the proof very briefly:

Suppose that the initial function ϕ depends only on a given linear combination of the space variables:

$\phi(x) = \psi(y)$, where y is the scalar quantity $x \cdot \xi$, ξ some (real) vector in the dual of x-space. We shall construct a solution u of (2.1) with initial value ψ which is a function of y and t only. Denote $u(x,t)$ by $v(y,t)$; we have the following equation for v:

$$v_t = Av_y \quad ,$$

where $A = \sum \xi_k A^k$.

To solve this equation, introduce new unknowns $w = Tv$, T so chosen that $\tilde{A} = TAT^{-1}$ is lower triangular, i.e. $\tilde{a}_{i\ell} = 0$ for $i < \ell$. According to known theorems of matrix theory, such a matrix T exists. It can be chosen to be real if the eigenvalues of A are real, and chosen so that the magnitude of T and T^{-1} (i.e. that of their elements) does not exceed a quantity that depends only on the dimension of the space. E.g., T can be chosen to be an orthogonal matrix.

The equation for w is

$$w_t = \tilde{A} w_y \quad . \tag{2.1'}$$

Since $\tilde{A}$ is lower triangular, the first equation contains w^1 only and has a unique solution once w^1 is specified on the y axis. The second equation contains w^1 and w^2 only; substituting the already known value of w^1 we can determine w^2 by integration along the second characteristic. Recursively we can determine all components and can give a bound for the magnitude of all components and their first derivatives in terms of bounds for the first $m+1$ derivatives of the data ψ independently of ξ.

Given an arbitrary but smooth function ϕ of x^1, ..., x^m, it can be represented as a superposition of "plane waves," i.e. of functions $\psi(x\,\xi)$ (see John, [13]). If ϕ is smooth, the functions ψ are likewise smooth, say $m+1$ times differentiable. It follows then from our previous estimate that the superposition of the corresponding solutions $v(x\cdot\xi,t)$ exists and has continuous first derivatives.

The much more delicate question of when Cauchy's problem is correctly posed for equations not homogeneous in the highest order term is discussed in Gårding, [7] and A. Lax, [14].

Take a point P outside of the initial hyperplane. It is well known in the theory of wave propagation that the value of u at P is uniquely determined by the initial data on a compact portion of the initial hyperplane (finite signal speed). We shall give now a precise description of this phenomenon:

The value of $u(P)$ is a linear functional depending continuously on the initial data ϕ in some C^k topology. Therefore $u(P)$ can be represented as a scalar product

$$u(P) = (\phi, R) \quad ,$$

where R is some distribution. It is well known in the theory of wave propagation that the support of R -- called the domain of dependence of P -- is compact.

This important property of solutions can be expressed without any reference to a representation as follows:

Theorem: Given any point P there exists a smallest closed set $K = K(P)$ on the initial hyperplane $t = 0$ such that if the initial values $\phi(x)$ of a solution u of (2.1) vanish in any open set O containing K, then $u(P) = 0$.

To prove this theorem we have to show that if $\{K\}$ is a collection of closed sets with the above property -- and from now on we shall call these sets of uniqueness -- then their intersection is also a set of uniqueness. We do this in two steps.

Lemma 2.1: The intersection of two sets of uniqueness is again a set of uniqueness.

Proof: We have to show that every solution u of (1.1), whose initial value $u(x,0) = \phi(x)$ is zero in an open set 0 containing $K_1 \cap K_2$, vanishes at P. Since the closed sets $K_1 \cap \overline{0}$ and $K_2 \cap \overline{0}$ are disjoint we can construct two open sets 0_1 and 0_2 containing $K_1 \cap \overline{0}$ and $K_2 \cap \overline{0}$ respectively, so that their closure $\overline{0}_1$ and $\overline{0}_2$ are also disjoint. Next we construct an infinitely differentiable function $p(x)$ which is equal to 0 in 0_1, equal to 1 in 0_2 and smooth everywhere. Define ϕ_1 and ϕ_2 as

$$\phi_1 = p\phi \quad , \qquad \phi_2 = (1-p)\phi \quad ,$$

and denote by u_1 and u_2 the solutions with initial values ϕ_1 and ϕ_2 respectively. The existence of u_1 and u_2 follows(2) from Theorem 2.1; furthermore, since solutions are uniquely determined by their initial data, it follows that the solution u with data ϕ is equal

---

2 At this point we use the hyperbolic character of (2.1).

to $u_1 + u_2$.

Clearly, ϕ_1 vanishes in $O_1 \cup O$ and ϕ_2 in $O_2 \cup O$; since each of these open sets contains a set of uniqueness (K_1 and K_2 respectively), it follows that $u_1(P)$ and $u_2(P)$ are zero; but then so is $u(P) = u_1(P) + u_2(P)$.

Repeated application of Lemma 2.1 shows that likewise the intersection of a finite number of sets of uniqueness is again a set of uniqueness.

Lemma 2.2: The intersection $K = \cap K_j$ of any number of sets of uniqueness K_j of which at least one is compact is again a set of uniqueness.

Proof: Let O be an open set containing $K = \cap K_j$. By the Heine-Borel Theorem, O also contains the intersection of a finite number of the sets K_j. Since a finite intersection was already shown to be a set of uniqueness, the conclusion follows.

Thus we have

Theorem 2.2: Every point P has a smallest set of uniqueness K on the initial hyperplane $t = 0$; this set is called the domain of dependence of the point P.

From now on we take the point P to be $(0,1)$, where 0 denotes the origin on the initial hyperplane. Its domain of dependence K is a pointset uniquely determined by the equation under consideration.

Whereas the precise determination of this pointset is a delicate matter, it is easy to determine its convex hull (i.e., the smallest convex set containing it) in terms of the coefficients A^k of equation (2.1).

We recall the notion of the support function $h(\xi)$ of the pointset K :

$$h(\xi) = \operatorname*{Sup}_{x \varepsilon K} x\xi \tag{2.3}$$

Here ξ is any (real) point of the dual of x-space and $x\xi$ denotes the scalar product $\sum x^j \xi_j$.

As is well known, the support function of a pointset is uniquely determined by its convex hull and vice versa.

Theorem 2.3: The support function of the domain of dependence K of the point P ; $(0,1)$ in equation (2.1) is

$$h(\xi) = \lambda_{max}\, A \cdot \xi \tag{2.4}$$

where $\lambda_{max}(X)$ denotes the largest eigenvalue of a matrix X and $A \cdot \xi$ abbreviates $\sum_j A^j \xi_j$.

Proof: We shall prove (2.4) for $\xi_1 = 1$, $\xi_2 = \xi_3 \ldots = \xi_m = 0$. We have to show that for any positive ε the halfspace $x^1 \leq \lambda_{max}(A^1) - \varepsilon$ is not a set of uniqueness whereas the halfspace $x^1 \leq \lambda_{max}(A^1) + \varepsilon$ is.

To show the first half of this proposition we have to exhibit a solution whose initial values vanish for $x^1 \leq \lambda_{max} - \varepsilon$ but which does not vanish at P. To construct such an example we introduce new unknowns $v = Tu$, T so chosen that $A = TA^1T^{-1}$ is lower triangular i.e. $a_{i\ell} = 0$ for $i < \ell$. For v, equation (2.1) reads

$$v_t = \sum TA^kT^{-1} v_{x^k} .$$

Our choice for v shall be a function that does not depend on the variables $x^2, \ldots, x^m$. So the equation for v reads

$$v_t = Av_{x^1} ,$$

where $A = TA^1T^{-1}$ is lower triangular. Componentwise, the equation reads

$$(2.1') \qquad v^i_t = \sum_\ell a_{i\ell}\, v^\ell_{x^1} .$$

The diagonal elements of A are the eigenvalues $\lambda_1, \ldots \lambda_n$ of A^1 in some order; say λ_{max} is the j^{th}:

$$a_{jj} = \lambda_{max} .$$

Choose the initial values of all but the jth component of v to be zero on the whole initial line, and choose $v^j(x,0)$ to be different from zero in some small neighborhood of $x^1 = \lambda_{max}$, and equal to zero outside this neighborhood. Solving the first $j-1$ equations in

(2.1$'$) recursively we conclude that the first $j-1$ components of v are zero for all x^1 and t. Therefore the j^{th} equation reads

$$v^j_t = \lambda_{max} \, v^j_{x^1} \quad ,$$

which asserts that v^j is constant along every line with slope $-\lambda_{max}$. In particular, $v^j(P)=v^j(0,1)=v^j(\lambda_n,0) \neq 0$.

To prove that the halfplane $x^1 \leq \lambda_{max} + \varepsilon$ is a set of uniqueness we rely on the

Holmgren Uniqueness Theorem: Let D be a domain in the hyperplane $t = 0$ and let $D(\tau)$, be a smooth deformation of D in x,t space keeping the boundary points fixed.[3] Suppose that none of the surfaces $D(\tau)$ is characteristic at any point; then D is a set of uniqueness for every point swept out by the family of surfaces.

We recall that a surface is characteristic at a point if the direction cosines n_x, n_t of its normal satisfy the characteristic relation

$$\det(n_t I - \sum n_{x^k} A^k) = 0$$

or, in other words, if n_t is an eigenvalue of

---

[3] I.e. a continuously differentiable mapping of the cylinder $\tilde{x} \in \bar{D}$, $0 \leq \tau \leq 1$ into x,t-space, $(\tilde{x},0)$ going into $x = \tilde{x}$, $t = 0$ and which is independent of τ on the boundary.

$A \cdot n$.

To apply Holmgren's Theorem to the situation at hand, choose D to be the domain which contains all points cut out of the half-space $x^1 \leq \lambda_{max} + \varepsilon$ by the sphere $|x| = R$.

Take the case that λ_{max} is positive; (if not the change of variables $\tilde{x}^1 = x^1 - at$ will accomplish this). We define the deformed surfaces $D(\tau)$ by

$$t = \tau(1 - \frac{x_1}{\lambda_{max}+\varepsilon})\,(1 - \frac{|x|}{R}) \;, \quad 0 \leq \tau \leq 1 \;.$$

For R large enough, $D(\tau)$ is nowhere characteristic for $0 \leq \tau \leq 1$. For we have

$$n_t = 1$$
$$n_{x^1} = \frac{\tau}{\lambda_{max}+\varepsilon} + O(\tfrac{1}{R})$$
$$n_{x^j} = O(\tfrac{1}{R}) \;, \qquad j = 2, \ldots, m \;.$$

Clearly, the matrix

$$I - \frac{\tau}{\lambda_{max}+\varepsilon} A^1$$

is nonsingular for ε positive and τ between zero and one, by definition of λ_{max} as the largest eigenvalue of A^1. But since the set of nonsingular matrices is open, the characteristic matrix is also nonsingular

$$I - A \cdot n$$

for R large enough.

Since the point $P = (0,1)$ lies on the surface $D(1)$, the proof of theorem 2.3 is complete.

Another proof of theorem 2.3 can be given on the basis of an explicit representation for the value of u at P in terms of the initial data. Such a formula can be derived as follows:

Define the "influence function" R as the solution of the backward initial value problem

$$L^* R = 0$$
$$R(x,1) = \delta(x)\, I \quad ,$$

where L^* denotes the adjoint of the operator L occurring in (2.1). This initial value problem can be solved explicitly: R is the Fourier transform of

$$e^{i(1-t)A^* \cdot \xi} \quad .$$

Apply Green's formula

$$\int_0^1 \int_{E^m} (R \cdot Lu - u \cdot L^* R)\, dx\, dt = \int_{E^m} u \cdot R\, dx \Big|_{t=0}^{t=1}$$

to any solution u of (2.1) and to the influence function R ; we obtain

$$u(P) = \int \phi(x)\, R(x,0)\, dx \quad .$$

The Fourier inverse of $R(x,0)$ is $e^{iA^* \cdot \xi}$; since we have assumed that all eigenvalues of $A \cdot \xi$ -- and thereby those of its adjoint, $A^* \cdot \xi$ -- are real for ξ real, the function $e^{iA^* \cdot \xi}$ grows at most polynomially for ξ real. For complex values of ξ, $e^{iA^* \cdot \xi}$ is entire analytic of exponential growth. Therefore, according to Plancherel and Polya's extension, see [17], of the Paley-Wiener theorem, the Fourier transform of $e^{iA^* \cdot \xi}$ has compact support K, and the support function $h(\eta)$ of the pointset K is equal to the rate of growth of $e^{A^* \cdot \eta}$. This brings us back to formula (2.4).

3. Some Matrix Theorems.

Theorem 3.1: *Let* $\{X\}$ *be a linear space of* n *by* n *matrices over the reals such that every* X *has only real eigenvalues. Then* $\lambda_{max}(X)$, *the largest eigenvalue, is a convex matrix function. Likewise* $\lambda_{min}(X)$, *the smallest eigenvalue, is a concave matrix function.*

Proof: Let $A^1, A^2, \ldots A^m$ be a generator of the space $\{X\}$. The differential equation (2.1) with these matrices as coefficients, is hyperbolic according to Theorem 2.1. Therefore, the concept of the domain of dependence $K(P)$, $P = (0,1)$, can be defined; according to Theorem 2.3, the support function $h(\xi)$ of $K(P)$, $P = (0,1)$ is $\lambda_n(X)$, $X = \sum \xi_j A^j$. But the support function of any domain, b being defined by (2.3) as the **supremum** of linear functions, is convex. This completes the proof.

In the symmetric case, this result follows from the classical maximum principle for the largest eigenvalue:

$$\lambda_{max}(X) = \sup_{\|a\|=1} (a, Xa) .$$

The concavity of $\lambda_{min}(X)$ follows from the relation $\lambda_{min}(X) = -\lambda_{max}(-X)$. From this it follows that the set of *positive (non-negative) matrices* X, i.e., matrices with positive (non-negative) eigenvalues, forms a convex cone. We introduce a partial ordering with respect to the cone of positive matrices:

Definition: *$A \le B$ means that all eigenvalues of $B - A$ are non-negative. Strict inequality is defined analogously.*

Since we have shown that over a linear space of matrices all of whose elements have only real eigenvalues the set of non-negative matrices forms a convex cone, it follows that over such a space this partial ordering is transitive, i.e. addition of inequalities and multiplication by positive scalars is permissible.

Theorem 3.2: *Let X and Y be two matrices all of whose linear combinations have real eigenvalues and suppose that X is negative. Let μ be an eigenvalue of $X + iY$:*

$$(X + iY)\, a = \mu a \quad . \tag{3.1}$$

Then the real part of μ is negative.

Proof: In the symmetric case the proof follows by taking the scalar product of (3.1) with the eigenvector a from which it follows that $\operatorname{Re}\mu = (a, Xa)$. In the general case consider the hyperbolic equation

$$u_t = Xu_x + Yu_y \quad ; \tag{3.2}$$

it follows from (3.1) that the function

$$u^{\tau}(x,y,t) = e^{(\mu t+x+iy)\tau}\, a \quad , \tag{3.3}$$

depending on the parameter τ, is a solution of (3.2) for all values of τ. Since by assumption $\lambda_{max}(X)$ is negative, it follows from formula (2.4) that the domain of dependence of the point $P = (0,1)$ is contained in the half-space $x < 0$. In this half-space the initial value of our function u^{τ} defined in (3.3) tends to zero as τ tends to plus infinity, and so do all its derivatives. Since according to Theorem 2.1 the solution is a continuous function of the data, $u^{\tau}(P) = e^{\mu\tau} a$ must also tend to zero as τ tends to plus infinity. This shows that the real part of μ is negative.

Likewise, if X is non-positive the real parts of the eigenvalues of $X + iY$ are non-positive.

Let X and Y be a pair of real matrices all whose linear combinations have only real eigenvalues. According to Theorem 3.1, the largest eigenvalue λ_{max} of $Y + sX$ is a convex, the lowest, λ_1, is a concave function of s. If in addition X is positive, both the smallest and largest eigenvalues of $Y + aX$ tend to $\pm \infty$ as s tends to $\pm \infty$. A convex or concave function with such a behavior at $\pm \infty$ is necessarily monotonic. Thus we have shown that the largest and smallest eigenvalues are monotonic matrix functions. We shall now prove a more general

Theorem 3.3: Let $\{X\}$ be a linear space of real matrices over the reals such that all eigenvalues of each X are real. Then the eigenvalues $\lambda_1(X), \lambda_2(X), \ldots, \lambda_n(X)$ are monotonic functions in the sense of the *partial ordering introduced before.*

For X and Y symmetric, the monotonicity of the eigenvalues follows from the classical min-max principle of Fisher, [5], see e.g. Riesz-Nagy, [19], p. 236, for the j^{th} eigenvalue:

$$\lambda_j = \underset{S_j}{\text{Min}} \ \underset{a \varepsilon S_j}{\text{Max}} \ \frac{(Xa,a)}{(a,a)} \ .$$

Proof: We have to show that the eigenvalues λ_j of $Y + sX, Y$ in $\{X\}$ and X positive, are monotonic functions of s. The eigenvalues $\lambda(s)$ of $Y + sX$ are branches of an algebraic function of s and so have only a finite number of singularities. We shall show that at every regular real point s_0 , $\lambda'(s_0)$ is non-negative. Then we can deduce that each branch of $\lambda(s)$ is non-decreasing in any interval of regularity. From this it follows that every branch of $\lambda(s)$ is strictly increasing; for otherwise λ would be constant in an interval and therefore, being analytic, would be constant for all s. But since X is positive, all eigenvalues of $Y + sX$ tend to infinity with s . Since λ is continuous at the possible branch points,(4) this

---

4 Actually, $\lambda(Y + sX)$ is regular on the whole real axis. For once, the proof of this result in the symmetric case due to Rellich, [18], can be carried out without alteration to the more general situation. Let s_0 be a singular point on the real axis; then $\lambda(s)$ has a Puiseux expansion in fractional powers of $(s-s_0)$. Since all branches of $\lambda(s)$ are real valued for s real, there can be no fractional, only integer, powers present in the expansion.

completes the proof of Theorem 3.3, except for showing that $\lambda'(s_0) \geq 0$ for regular real s_0. Assume now to the contrary that at a regular real point s_0 -- which to simplify the notation, we take to be zero -- λ' is negative, say $\lambda'(0) = -a$, $a > 0$. Then we have, expanding about $s = 0$,

$$\lambda(Y+sX) = \lambda(Y) - as + O(|s|^2) .$$

Since $s = 0$ is a regular point, the above relation holds for complex values of s as well. Put $s = it$ and multiply the relation by i; we get

$$\lambda(-tX+iY) = i\lambda(Y) + at + O(t^2) .$$

This shows that for t positive and small enough the matrix $-tX + iY$ has an eigenvalue with positive real part. But since X was assumed positive, this contradicts Theorem 3.2.

We indicate now another possible approach to the results presented in this section:

Let $p(\xi, \eta, \lambda)$ be a form of degree n in ξ, η, λ, with the following property: The coefficient of λ^n is one, and for each fixed real choice for ξ and η, $p(\xi, \eta, \lambda)$ has only real zeros in λ. If A and B are a pair of symmetric n by n matrices, the form

$$p(\xi, \eta, \lambda) = \det |\xi A + \eta B - \lambda I|$$

has this property. Conversely, I conjecture that every form with the above property can be so represented. In support of this conjecture is the fact that the number of coefficients in p -- not counting that of λ^n -- is equal to the number of elements of A and B if one of these matrices is normalized by an appropriate orthogonal transformation to be diagonal. If true, this representation theorem could be used to deduce the convexity and monotonicity of eigenvalues over a linear space of matrices with all real eigenvalues from the classical case of symmetric matrices.

A similar conjecture has been made by Chandler Davis (personal communication).

4. Finite Difference Schemes.

In this section we consider one level explicit difference equations approximating the differential equation (2.1), i.e. difference equations of the form

$$(4.1) \qquad u(x,t+\Delta t) = \sum_j c^j\, u\,(x+\Delta_j,t) \;.$$

This equation expresses the value of u at the point x and at the time $t+\Delta t$ as a linear combination of values of u at time t and at points $x+\Delta_j$. The displacements Δ_j will be taken to depend linearly on Δt :

$$(4.2) \qquad \Delta_j = \Delta t\; r_j \;;$$

the vectors r_j remain fixed as Δt tends to zero.

Such a difference equation is said to be derived from, or consistent with, the differential equation (2.1) if every smooth solution of (2.1) satisifies (4.1) approximately, i.e. if the two sides differ by an amount $o(\Delta t)$. Expand $u(t+\Delta t)$ and $u(x+\Delta_j)$ into a Taylor series; the left and right sides of (4.1) can then be written in the form

$$(4.1') \qquad u + \Delta t\; u_t + O(\Delta^2)$$

and

$$(4.1'') \qquad \sum c^j\, u + \sum c^j\, u_x \cdot \Delta_j + O(\Delta^2) \;.$$

Substitute $\Delta \cdot u_x$ for u_t as given by (2.1) into (4.1') and $\Delta t r_j$ for Δ_j into (4.1'') according to (4.2). The two differential operators (4.1') and (4.1'') have to agree up to terms of order Δ^2, which leads to the relations

$$\sum C^j = I \tag{4.3}$$

$$\sum C^j r_j = A \quad . \tag{4.4}$$

The initial value $u(x,0) = \phi(x)$ uniquely determines a solution of the difference equation (4.1) at all times which are integer multiples of Δt. The scheme based on (4.1) is said to be convergent if the solutions u_Δ of the difference equation converge uniformly as Δt tends to zero to the solution u of the differential equation (2.1) with the same initial value ϕ, for all sufficiently differentiable data ϕ.

Suppose the initial function is an exponential:

$$\phi(x) = e^{ix\xi} v \;, \qquad v = \text{const.}$$

The corresponding solution of (4.1) is also an exponential, given by the formula

$$u(x,t) = e^{ix\xi} \left(\sum C^j e^{i\Delta_j \cdot \xi}\right)^N v \quad , \tag{4.5}$$

where $t = N\Delta t$. The matrix $\sum C^j e^{ir_j\zeta}$ is called the amplification matrix or generating function, and will be denoted by $G(\zeta)$.

Theorem 4.1: The necessary and sufficient condition for the convergence of a difference scheme is that the eigenvalues of the associated amplification matrix $G(\zeta)$ should not exceed one for real ζ .

This condition is due to v . Neumann (see e.g. [4], [11] or [15]).

The necessity of this condition, as already observed in [15] within the framework of L_2 convergence, follows from the principle of uniform boundedness (see [1]). For let P be any point not on the initial hyperplane, say $P = (0,1)$; $u_\triangle(P)$ is a bounded linear functional of the initial value ϕ ;

$$u_\triangle (P) = L_\triangle \phi \quad .$$

If $u_\triangle(P)$ converges to $u(P)$ for every ϕ in the space C^N of N times differentiable functions, then a fortiori the functionals $L_\triangle \phi$ remain bounded for every ϕ in C^N , and so according to the principle of uniform boundedness, the functionals L are uniformly bounded.

Suppose now that v. Neumann's condition is violated, i.e. for some real value of $\zeta, G(\zeta)$ has an eigenvalue λ of absolute value greater than one:

$$G(\zeta)v = \lambda v \ , \qquad |\lambda| > 1 \quad .$$

Consider the set of data

$$\phi_{\Delta} = e^{i\zeta/\Delta t} v \quad .$$

According to (4.5), $L_{\Delta}\phi_{\Delta} = u_{\Delta}(P)$ is equal to

$$L_{\Delta}\phi_{\Delta} = G(\zeta)^N v = \lambda^{1/\Delta t} \; e^{i\zeta/\Delta t} \; v \quad .$$

As Δt tends to zero, the C^N norm of ϕ_{Δ} grows like $(1/\Delta t)^N$, while $|L_{\Delta}\phi_{\Delta}|$ tends to infinity exponentially in $1/\Delta t$; this shows that $|\lambda| > 1$ precludes the uniform boundedness of the functionals L_{Δ}.

The sufficiency of the condition $|\lambda| \leq 1$ for convergence will be shown by expanding the given data ϕ in a Fourier series. This is a classical procedure; since the proof usually is carried out under more restrictive conditions, we shall sketch the details here:

Lemma 4.1: Under the assumption of theorem 4.1, the powers of the amplification matrix G satisfy the inequality[5]

$$\|G^N(\zeta)\| \leq \text{const.} + (N|\zeta|)^{n-1} \quad .$$

Proof: Write G in the form

---

[5] By the norm of a matrix we mean the absolute value of its largest component.

$$G = \sum C^j + \sum C^j \left(e^{ir_j \cdot \zeta} - 1\right) ;$$

according to (4.3) this can be written as

$$G = I + \sum C^j \left(e^{ir_j \cdot \zeta} - 1\right) = I + H \quad . \tag{4.6}$$

Clearly,

$$\|H\| = O(\zeta) \quad .$$

Let S be a matrix such that

$$S^{-1}HS = H'$$

is triangular. Such an S exists, and it can be so chosen that the norms of S and of S^{-1} are bounded by a quantity that depends only on the order n of the matrices in question. Therefore the size of H' does not exceed a constant multiple of $\|H\|$;

$$\|H'\| \leq \text{const} \|H\| = O(|\zeta|) \quad . \tag{4.7}$$

From (4.6) we have:

$$G' = S^{-1}GS = I + H' ; \tag{4.6'}$$

G' also is a triangular matrix. Separate G' into a diagonal and subdiagonal part:

$$G' = D + T \quad . \tag{4.8}$$

Since T is the subdiagonal part of H' its norm is bounded by that of H ; so we have

$$(4.7') \qquad \|T\| = 0\,(|\zeta|) \quad .$$

The diagonal elements of D are the eigenvalues of G' ; these are the same as the eigenvalues of G which, by the assumption of Theorem 4.1, do not exceed 1 in absolute value.

Raise both sides of (4.8) to the Nth power; we obtain a representation for $(G')^N$ as a sum of products of N factors D and T. Since D is diagonal and T is subdiagonal, all terms which include n or more factors of T are zero. Since furthermore the eigenvalues of D do not exceed one in absolute value and since the norm of T is bounded by (4.7'), we get the following estimate for $(G')^N$.

$$\|(G')^N\| \leq \text{const.} + (N|\zeta|)^{n-1} \quad .$$

Since G^N is equal to $S(G')^N S^{-1}$, a similar estimate holds for G^N ; this proves Lemma 4.1 .

Lemma 4.2: Under the hypothesis of Theorem 4.1, the linear functionals L_Δ relating the value of solutions of (4.1) at $P = (0,1)$ to their initial values ϕ are uniformly bounded for all values of Δ in the C^M norm for M sufficiently large.

Proof: Represent ϕ in C^M as a Fourier transform(6):

---

(6) We assume that ϕ is zero outside of an open set containing the domain of dependence of P .

$$(4.9) \qquad \phi(x) = \int \psi(\xi) e^{ix\xi} d\xi \quad .$$

The solution $u_\triangle$ of (4.1) with initial value ϕ given by (4.9) can be written in the form

$$(4.10) \quad u_\triangle(x,t) = \int G^N(\triangle t\xi)\psi(\xi)e^{ix\xi}d\xi \quad , \quad t = N\triangle t \quad .$$

This formula is obtained by superposition from formula (4.5). To prove its validity we have to show that the integral on the right in (4.10) converges. To this end we use the estimate of Lemma (4.1) which yields the following majorant for the integral (4.10):

$$\int \left\{ \text{const} + (N\triangle t|\xi|)^{n-1} \right\} |\psi(\xi)| d\xi \quad ;$$

at $t = N\triangle t = 1$ this integral is equal to

$$\int \left\{ \text{const} + |\xi|^{n-1} \right\} |\psi(\xi)| d\xi \quad .$$

But this last integral converges if ψ is the Fourier inverse of a sufficiently differentiable function ϕ. The order k of the required differentiability depends only on n.

This reasoning shows at the same time that the value of $|u_\triangle|$ at P is bounded by a constant multiple of the C^M norm of ϕ, for M large enough. This completes the proof of Lemma 4.2.

Next we verify that for an exponential initial function $\phi = e^{ix\xi}v$, the corresponding solution $u_\triangle$

of the difference equation (4.1) does tend to the exact solution u of the differential equation (2.1). According to (4.5), $u_{\Delta}(x,t)$ is given by the formula

$$(4.11) \qquad u_{\Delta}(x,t) = G^N(\Delta t\xi)e^{ix\xi}v \ , \quad N\Delta t = t \ .$$

Using the definition of G and relation (4.3), we can write $G(\Delta t\xi)$ as

$$\begin{aligned} G(\Delta t\xi) &= I + \sum C^j(e^{i\Delta_j\xi} - 1) \\ &= I + i\sum C^j \Delta_j\xi + O(\Delta t^2) \\ &= I + i\Delta t \sum C^j r_j \cdot \xi + O(\Delta t^2) \ . \end{aligned}$$

Raise both sides to the Nth power, $N = t/\Delta$, and take the limit as Δt tends to zero. Using the classical rule

$$\lim_{\Delta t \to 0} (I + \Delta t\, a + O(\Delta t^2))^{t/\Delta t} = e^{at} \ ,$$

we obtain for the limit on the right in (4.1) the value

$$e^{i[t\sum C^j r_j \cdot \xi + x\xi]} v \ .$$

According to the relation (4.4) between the coefficients of the difference and differential equation, this is equal to

$$e^{i[tA \cdot \xi + x\xi]} v$$

which is, it is easy to check, an exact solution of the differential equation (2.1) taking on the assigned initial values.

This shows that the solution of the difference equation with prescribed exponential initial data tends to the corresponding solution of the differential equation. By the principle of superposition the same is true for a solution with finite linear combinations of exponentials as initial data. Since the exponential functions span the space C^k, and since furthermore the functionals L_Δ were shown in Lemma 4.2 to be uniformly bounded, it follows by a standard argument (Banach-Steinhaus) that $L_\Delta \phi$ converges to $u(P)$ for all ϕ in C^k. This completes the proof of Theorem 4.1.

Theorem 4.1 is an exact analogue of Theorem 2.1; v. Neumann's condition plays the same role for difference equations as the condition of hyperbolicity for differential equations.

If we want to use v. Neumann's condition to verify that some proposed difference scheme converges, we have to check that the eigenvalues of the amplification matrix $G(\zeta)$ do not exceed one in absolute value for all real values of ζ. This is not always an easy task, as any numerical analyst who has tried a fairly complicated case can testify. Therefore there is room for more directly verifiable sufficient conditions for convergence. Such a one has been given by Friedrichs, see [6] or [15] for symmetric difference schemes; our aim in this section is to extend it to not necessarily symmetric schemes.

Theorem 4.2: If all linear combinations with real coefficients of the coefficients C^j of the difference equation (4.1) have only real eigenvalues, and if each C^j is non-negative, then the scheme based on (4.1) is convergent.

We shall show that under the assumptions of theorem 4.2 all matrices of the form

$$\sum z_j C^j \quad ,$$

z_j complex numbers of modulus one, have eigenvalues not exceeding one in modulus. Since the amplification matrix G is of this form, with $z_j = e^{ir_j \cdot \xi}$, this shows that v. Neumann's condition is satisfied.

The symmetric case is easy to handle; let λ be an eigenvalue, v the corresponding normalized eigenvector:

$$\lambda v = \sum z_j C^j v \quad .$$

Take the scalar product of both sides with v :

$$\lambda(v,v) = \sum z_j (C^j v, v) \quad .$$

The absolute value of the left side is $|\lambda|$, while that of the right does not exceed $\sum (C^j v, v)$; here we have used the fact that C^j is non-negative. But according to (4.3), this last expression is equal to $(v,v) = 1$.

We turn now to the nonsymmetric case. Consider the matrix

$$\sum z_j C^j - I \quad ;$$

and replace I by $\sum C^j$, as in (4.3), and separate real and imaginary parts. We get

$$\sum (x_j - 1)\, C^j + i \sum G_j\, C^j$$

where x_j and y_j are the real and imaginary parts of z_j. Since x_j cannot exceed 1, the real part of the above matrix is a linear combination of non-negative matrices with nonpositive coefficients and is therefore by our convexity theorem, itself non-positive. Then, according to theorem (3.2), the real part of the eigenvalues of $\sum z_j C^j - I$ are also non-positive. So we have shown:

The real parts of the eigenvalues of $\sum z_j C^j$ *do not exceed one.* But since the eigenvalues of $\sum z_j C^j$ are homogeneous functions of degree one, this shows that the moduli of the eigenvalues cannot exceed 1 either.

It is easy to show that the conditions in theorem 4.2 on C^j are not only sufficient but necessary for $\sum z_j C^j$ to have all its eigenvalues in the unit circle for all possible choices of z_j, $|z_j| = 1$. In case that the vectors r_j do not lie on a lattice, it can be shown that v. Neumann's condition implies that $\sum z_j C^j$ has no eigenvalues of modulus greater than one for all possible choice of z_j, $|z_j| = 1$ (see [9] for a proof). Therefore in this case the conditions contained in theorem 4.2 are necessary as well as sufficient for the convergence of a difference scheme.

In most difference schemes the coefficients C^j are linear combinations of the coefficients of the hyperbolic differential equation which the scheme is supposed to approximate. In such a situation the first assumption of Theorem 4.2 -- about the reality of the eigenvalues -- is automatically fulfilled, since the coefficients of the given hyperbolic equation have this property.

In the symmetric case Friedrichs has proved more; he has shown that under the hypothesis of Theorem 4.2 the difference operator (4.1) relating the value of u at t to its value at $t + \Delta t$ does not increase the L_2 norm with respect to the space variables:

$$\int u^2(x,t+\Delta t)dx \leq \int u^2(x,t)dx \quad .$$

Furthermore he has shown that even if the coefficients C^j depend on the variables x and t, we have

$$\int u^2(x,t+\Delta t)dx \leq (1+O(\Delta t)) \int u^2(x,t)dx \quad ,$$

provided that C^j are Lipschitz continuous. This guarantees, it is easy to show, the convergence of the difference scheme.

We shall now present some applications of Theorem 4.2.

i) Convert the differential equation (2.1) into a difference equation by replacing the x^k derivative by a symmetric difference quotient:

$$u_{x^k} \sim \frac{u(x+\delta_k e_k) - u(x-\delta_k e_k)}{2\delta_k}$$

where e_j is the unit vector in the x^j direction, and by replacing the time derivative by a forward difference quotient:

$$u_t \sim \frac{u(x,t+\Delta t) - \bar{u}(x,t)}{\Delta t} \quad ,$$

where $\bar{u}$ is the mean value of u at the $2m$ points $x \pm \delta_k e_k$:

$$\bar{u}(x,t) = \frac{\sum_{k=1}^{m} \left\{ u(x+\delta_k e_k, t) + u(x-\delta_k e_k, t) \right\}}{2m}$$

The resulting difference scheme is of the form (4.1), with the $2m$ displacement vectors $\pm \delta_k e_k$, and with coefficients

$$C_{\underline{+}k} = \frac{1}{2m} I \pm \frac{\Delta t}{2\delta_k} A^k \quad .$$

Applying Theorem 4.2 shows that this difference scheme is stable if Δt is chosen so small that $\delta_k/m\Delta t$ exceeds the absolutely largest eigenvalue of A^k . This is am improvement of a result due to Fritz John who has shown (in an unpublished note) that the above difference scheme is stable if Δt is chosen to be small enough.

ii) Susan Hahn, in [9], has investigated simplicial difference schemes, i.e., schemes in which the number of displacement vectors r_j is one greater than the number of space variables x and in which the displacement vectors are the vertices of a simplex. She has shown that if such a scheme satisfies the classical necessary condition for convergence due to Courant, Friedrichs and Lewy, [2], i.e., if the domain of dependence of the point $P = (0,1)$ lies in the simplex formed by the displacement vectors r_j , then the coefficients C_j are non-negative. Thus she was able to conclude -- using Friedrichs' theorem in the symmetric case and Theorem 4.2 in the general case -- that for a simplicial scheme the Courant-Friedrichs-Lewy condition is sufficient as well as necessary for convergence.

BIBLIOGRAPHY

[1] Banach, S., Opérations Lineaires, 1932. Warsaw.

[2] Courant, R., Friedrichs, K. O., and Lewy, H., *Über die partielle Differenzialgleichurgen der mathematischen Physik*, Math. Annalen, Vol. 100, 1928, pp. 32-74.

[3] Courant, R. and Lax, A., *Remarks on Cauchy's Problem*, Comm. on Pure and Applied Mathematics, Vol. 8, 1955, pp. 497-502.

[4] Eddy, R. P., *Stability in the numerical solution of initial value problems in partial differential equations*, Naval Ordnance Lab. Memo. 10232.

[5] Fisher, E., *Über quadratische Formen mit reellen Koefficienten*, Monatshefte fur Math. und Phys., Vol. 16 (1905), pp. 234-249.

[6] Friedrichs, K. O., *Symmetric Hyperbolic Linear Differential Equations*, Comm. Pure and Applied Mathematics, Vol. 7, 1954, pp. 345-392.

[7] Gårding, L., *Linear hyperbolic partial differential equations with constant coefficients*, Acta Math., Vol. 85, 1951, pp. 2-62.

[8] Hadamard, J., *Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques* (Herman, 1932).

[9] Hahn, S., *Dissertation*, New York University, 1957.

[10] Holmgren, E., *Über Systeme von linearen partiellen Differential gleichurgen*, Öfversigt Kongl. Vetens.-Akad. Förh., Vol. 58, 1901, pp. 91-103.

[11] Hyman, M. A., O'Brien, G. G., Kaplan, S., *A Study of the numerical solutions of partial differential equations*, Journal of Mathematical Physics, Vol. 29, 1951, pp. 223-251.

[12] John, F., *On linear partial differential equations with constant coefficients*, Comm. on Pure and Applied Mathematics, Vol. 2, 1949, pp. 209-253.

[13] John, Fritz, *Plane Waves and Spherical Means*, New York, Interscience, 1956.

[14] Lax, A., On Cauchy's problem for partial differential equation with multiple characteristics, Comm. Pure and Appl. Math., Vol. 9, 1956, pp. 135-168.

[15] Lax, P. D., Richtmyer, R. D., Survey of the stability of linear finite difference equations, Comm. Pure and Appl. Math., Vol. 9, 1956, pp. 267-293.

[16] Lax, P. D., Asymptotic solutions of oscillatory initial value problems, Duke Math. Journal, to appear in December 1957.

[17] Plancherel, M., Polya, G., Fonctions entieres et integrales de Fourier multiple, Commentarii Matematici Helretici, Vol. 9, 1936-37, pp. 224-248.

[18] Rellich, F., Störungstheorie der Spectralzerlegung, Math. Annelen, Vol. 113 (1936), pp. 600-619.

[19] Riesz, F., Nagy, Sz.-B., Leçons d'analyse fonctionnelle, Akadémiai Kiadó, Budapest, 1952.

Date Due

JAN 30 '58	APR 12 1977		
FEB 13 '59	FEB 1		
MAR 24 '59			
OCT 14 '59	AUG 2		
DEC 7 '59	SEP 11 1980		
FEB 28 '60	OCT 2 1980		
MAR 9 '60			
SEP 6 '60			
DEC 17 '65			
JAN 5 '66			
[illegible] '66			
OCT 5 '67 NOV 11 '68			
MAR 23 '69			
MAR 6 1976 FEB 2			
OCT 28 1976	PRINTED	IN U. S. A.	

NYU
NYO
7974

Lax
Differential equations,
Difference equations and
matrix theory.

NYU NYO 7974

Lax
AUTHOR

Differential equations,
TITLE difference equations, and
matrix theory.

DATE DUE	BORROWER'S NAME
~~APR 1 2 1977~~	~~Moon~~
~~FEB 1 8 1978~~	~~I. C. CHANG~~
[illegible]	[illegible]

N. Y. U. Institute of
Mathematical Sciences
25 Waverly Place
New York 3, N. Y.